Adaption of Akaike Information Criterion under Least Squares Frameworks for Comparison of Stochastic Models

H.T. Banks and Michele L. Joyner

Center for Research in Scientific Computation
North Carolina State University
Raleigh, NC 27695
and
Dept of Mathematics and Statistics
East Tennessee State University
Johnson City, TN 37614

April 16, 2019

Abstract

In this paper, we examine the feasibility of extending the Akaike Information Criterion (AIC) for deterministic systems as a potential model selection criteria for stochastic models. We discuss the implementation method for three different classes of stochastic models: continuous time Markov chains (CTMC), stochastic differential equations (SDE), and random differential equations (RDE). The effectiveness and limitations of implementing the AIC for comparison of stochastic models is demonstrated using simulated data from the three types of models and then applied to experimental longitudinal growth data for algae.

Key Words:continuous time Markov chain models, CTMC, stochastic differential equations, SDE, random differential equations, RDE, inverse problems, model comparison techniques, Akaike Information Criterion, AIC

Mathematics Subject Classification:93E03, 37L55, 37H10, 94A30

1 Introduction

Developing a mathematical model to describe a physical system involves an iterative process [13] which includes 1) the formalization of properties and relationships within the system, 2) the abstraction of those concepts into mathematical relationships, 3) the inclusion and formalization of uncertainty in the model, 4) the analysis of the numerical solution, 4) the comparison of the output of the model to the data or real system, 5) the incorporation of changes in the model based on this comparison and then 6) the cycle starts again. One hidden step in this process is that there may be multiple possibilities for how to describe the relationships or processes within the system. For example, consider a structured population model for *Daphnia magna*, a modern day "canary in the mine shaft" in ecology. In developing a model for *D. magna* [1, 25], data was collected in the laboratory where the daphnia were fed green algae. In this modeling process, it was necessary to incorporate the growth of the green algae into the structured population model for *Daphnia magna*. However, how does one determine which growth model (logistic, Bernoulli, Gompertz, etc.) best describes the growth of green algae? This is accomplished using model comparison techniques. Model comparison techniques have been widely developed for deterministic models [7, 10, 14, 15, 16, 22]; however, they are still in their infancy for stochastic models. In a paper by Banks and Joyner [11], an extension of the residual sum of squares technique for model comparison was successfully used to compare two nested stochastic models. In this paper, we seek to extend the Akaike Information Criterion (AIC) [7, 14, 15, 16, 22] for stochastic models which are *not* nested using the conceptual basis presented in [11].

Two models are considered nested when one model can be obtained from the other by assuming a restricted parameter space $\Omega_q^H \subset \Omega_q$ such that $\Omega_q^H = \{\mathbf{q} \in \Omega_q : \mathcal{H}\mathbf{q} = \mathbf{h}\}$. For example, the logistic growth model,

$$\frac{dx}{dt} = rx(t)\left(1 - \frac{x(t)}{\kappa}\right), \tag{1}$$

and Bernoulli growth model,

$$\frac{dx}{dt} = rx(t)\left(1 - \left(\frac{x(t)}{\kappa}\right)^{\beta}\right), \tag{2}$$

are considered nested models, because the logistic model can be obtained from the Bernoulli model when the parameter β in the Bernoulli model is set equal to 1. In both of these models, the parameter r describes the growth rate and κ the limiting capacity. The Gompertz growth model,

$$\frac{dx}{dt} = \alpha x(t)\left(\log\kappa - \log x(t)\right) = \alpha x(t)\log\left(\frac{\kappa}{x(t)}\right), \tag{3}$$

is another growth model where growth is slowest at the beginning and when the population reaches its limiting capacity, κ. Here α denotes a scaling parameter. Although the Gompertz model is related to the logistic model through a limiting process [25], it is not a nested model with either the logistic or Bernoulli growth models. In other words, one model cannot be obtained from the other by restricting the parameter space. If one wanted to compare a stochastic Gompertz growth model to a stochastic logistic or Bernoulli growth model, one could not use the techniques developed in [11] since the models aren't nested. Therefore, in this paper, we focus on a model comparison technique for stochastic models which are **not** nested. In particular, we concentrate on the continuous-time Markov chain (CTMC), stochastic differential equation (SDE) and random differential equation (RDE) models for multiple growth models. The stochastic models we use for validation purposes are given in Section 2. In Section 3, we summarize how one can approximate a stochastic model with a deterministic model as in [11], and then, in Section 4, we summarize how one can then use the Akaike Information Criterion for model comparison of the various stochastic models. We give

the results using synthetic data from CTMC, RDE and SDE growth models in Section 5 and discuss the benefits and limitations with this method. Finally, in Section 6, we apply this technique to experimental longitudinal data for algae and then conclude the paper with some final remarks in Section 7.

2 Models

Deterministic models, like the growth models in Equations (1)-(3), incorporate no randomness; therefore, given the same set of parameters and initial conditions, the output of a deterministic model will always produce the same unique trajectory. Continuous-time Markov chain (CTMC), stochastic differential equation (SDE) and random differential equation (RDE) models all add randomness in the model but in different ways and thus result in infinitely many possible trajectories. A CTMC model satisfies the Markov property, or memoryless property, which means that the next state depends only on the value of the current state and not on the history of the process [3]. CTMC models are formulated as the probability of a transition from one state to all others and, therefore, the change in state no longer occurs with certainty. An Itô SDE model also incorporates randomness into the model, but in a different manner. The general form for an Itô SDE can be derived from a CTMC model and is given by

$$dX(t) = \mu(t, X(t))dt + B(t, X(t))dW(t), \quad t \geq 0, \tag{4}$$

where

$$\mu(t, X) = \frac{E(\Delta X)}{\Delta t}, \quad B(t, X) = V^{1/2} \text{ with } V = \frac{E(\Delta X \Delta X^T)}{\Delta t},$$

and W is a Wiener process such that $W(0) = 0$ and

$$W(t) - W(s) \approx \mathcal{N}(0, t - s).$$

In a CTMC model, there are infinitely many different trajectories due to the different transition probabilities. In the SDE model, there are still an infinite number of trajectories possible; however, in the SDE model, it is due to the Wiener process term. In a RDE model, the randomness is included through the parameters. A general random ordinary differential equation (RDE) containing random parameter values can be written as

$$\frac{d\mathbf{x}}{dt} = g(t, \mathbf{x}, \mathbf{Q}), \quad \mathbf{x}(0) = \mathbf{x}_0 \tag{5}$$

where $\mathbf{Q}$ is a m-dimensional random vector. For example, one may formulate a logistic RDE model by assuming that one or both of the parameter values in the model (r or κ) are random variables which behave according to some known distribution. For example, let $R \sim \mathcal{N}(\mu_R, \sigma_R^2)$ and assume κ is constant, then

$$\frac{dx(t;Q)}{dt} = Rx(t;Q)\left(1 - \frac{x(t;Q)}{\kappa}\right) \tag{6}$$

is a random differential equation with random variable parameter $Q = R$. The growth models we consider in this paper are the logistic growth model, Bernoulli or generalized logistic growth model, the Gompertz model, dynamic carrying capacity model, power law model, von Bertalannfy model, exponential growth model and exponential-linear growth model. The deterministic systems are given in Table 1 together with the total number of parameters in the system with and without assuming the initial condition is unknown. A comparison of the different trajectories for a sample of parameters is shown in Figure 1. We note that the number of parameters is important when using the AIC model comparison technique as increased complexity results in a penalization term in AIC formulation. We give each type of stochastic growth model in Sections 2.1 - 2.8.

Table 1: **Deterministic Growth Models for Model Comparison**

Model	Equation	No. Param w/o IC	No. Param w/ IC
Logistic	$\frac{dx}{dt} = rx\left(1 - \frac{x}{\kappa}\right)$ $x(0) = x_0$	2	3
Bernoulli (Generalized Logistic)	$\frac{dx}{dt} = rx\left(1 - \left(\frac{x}{\kappa}\right)^{\beta}\right)$ $x(0) = x_0$	3	4
Gompertz	$\frac{dx}{dt} = ax\log\left(\frac{\kappa}{x}\right)$ $x(0) = x_0$	2	3
Dynamic Carrying Capacity	$\frac{dx}{dt} = ax\log\left(\frac{\kappa}{x}\right)$ $\frac{d\kappa}{dt} = bx^{2/3}$ $x(0) = x_0;\ \kappa(0) = \kappa_0$	2	4
Power Law	$\frac{dx}{dt} = ax^{\mu}$ $x(0) = x_0$	2	3
von Bertalanffy	$\frac{dx}{dt} = ax^{\mu} - bx$ $x(0) = x_0$	3	4
Exponential	$\frac{dx}{dt} = ax$ $x(0) = x_0$	1	2
Exponential Linear	$\frac{dx}{dt} = a_0 x,\ t \le \tau$ $\frac{dx}{dt} = a_1,\ t > \tau$ $x(0) = x_0;\ \tau = \frac{1}{a_0}\ln\left(\frac{a_1}{a_0 x_0}\right)$	2	3

2.1 Logistic Stochastic Models

The deterministic logistic model was given above in Equation (1). There are multiple different birth-death CTMC models [3] for which the limiting deterministic model is given by Equation (1). In this paper, we only consider the CTMC model given by

$$\text{Prob}(\Delta X = j | X(t) = x) = \begin{cases} rx\Delta t + o(\Delta t) & j = 1 \\ \frac{rx^2}{\kappa}\Delta t + o(\Delta t) & j = -1 \\ 1 - \left(rx + \frac{rx^2}{\kappa}\right)\Delta t + o(\Delta t) & j = 0 \\ o(\Delta t) & j \neq 1, -1, 0. \end{cases} \tag{7}$$

We note that the differences exhibited in realizations for two different CTMC models limiting to the same deterministic model can be substantial when considering small population sizes [3]; however, for the population sizes we consider in this work, the differences in the realizations are not substantial (see [3] and [11] for examples).

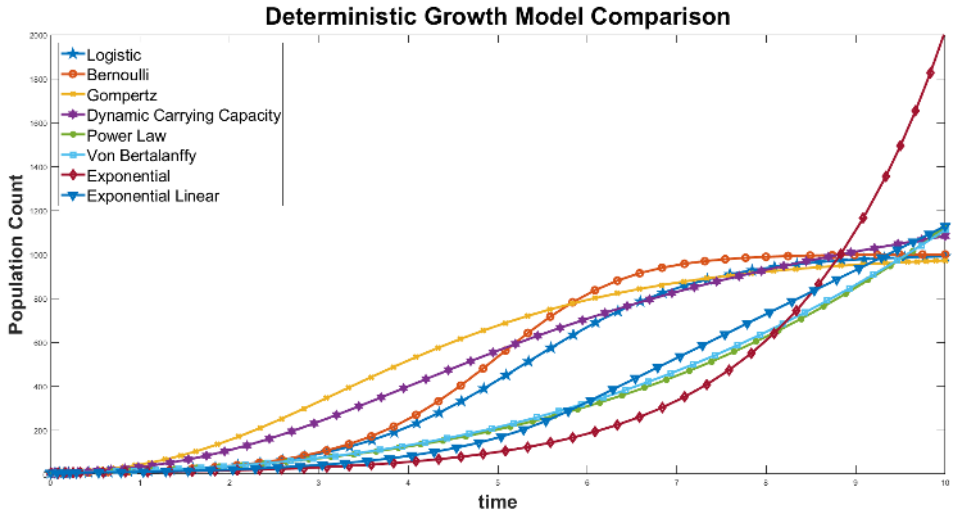

Figure 1: This figure shows the various trajectories for the nine models given in Table 1 with model parameters given by: Logistic ($r = 1$, $\kappa = 1000$), Bernoulli ($r = 1$, $\kappa = 1000$, $\beta = 1.5$), Gompertz ($a = 1.2$, $\kappa = 1000$), Dynamic Carrying Capacity ($a = 1$, $b = .4$), Power Law ($a = 2.2$, $\mu = 0.7$), von Bertalanffy ($a = 2.2$, $\mu = 0.8$, $b = 0.3$), Exponential ($a = 0.6$), and Exponential-Linear ($a_0 = 0.7$, $a_1 = 200$).

Using the technique outlined in [3], the Itô SDE model equivalent to the logistic CTMC model in Equation (7) is given by

$$dX(t) = rX(t)\left(1 - \frac{X(t)}{\kappa}\right) + \sqrt{rX(t)\left(1 + \frac{X(t)}{\kappa}\right)}\ dW. \tag{8}$$

As explained above, a random differential equation model is determined by assuming one or more parameters in the deterministic model are random variables. Equation (6) is one example where the growth rate is the only random variable parameter. One can compare the consequence of assuming fixed versus random variables in [12]; however, in generating synthetic data, we only assume the growth rate $R \sim \mathcal{N}(\mu_R, \sigma_R^2)$ is a random parameter as in Equation (6).

2.2 Bernoulli Stochastic Models

The deterministic Bernoulli growth model is described previously and given by Equation (2). As mentioned, although there are multiple CTMC models which have an equivalent deterministic approximation, we only consider the CTMC Bernoulli model given by

$$\text{Prob}(\Delta X = j | X(t) = x) = \begin{cases} rx\Delta t + o(\Delta t) & j = 1 \\ rx\left(\frac{x}{\kappa}\right)^{\beta}\Delta t + o(\Delta t) & j = -1 \\ 1 - \left(rx + rx\left(\frac{x}{\kappa}\right)^{\beta}\right)\Delta t + o(\Delta t) & j = 0 \\ o(\Delta t) & j \neq 1, -1, 0. \end{cases} \tag{9}$$

An Itô SDE model can be derived from the CTMC model in Equation (9) and is given by

$$dX(t) = rX(t)\left(1 - \left(\frac{X(t)}{\kappa}\right)^{\beta}\right) + \sqrt{rX(t)\left(1 + \left(\frac{X(t)}{\kappa}\right)^{\beta}\right)}\ dW. \tag{10}$$

One RDE Bernoulli growth models can be found by assuming the growth parameter R is a random variable, $R \sim \mathcal{N}(\mu_r, \sigma_R^2)$, and κ is constant,

$$\frac{dx(t;Q)}{dt} = Rx(t;Q)\left(1 - \left(\frac{x(t;Q)}{\kappa}\right)^{\beta}\right). \tag{11}$$

Note that the logistic stochastic growth models can be obtained by setting $\beta = 1$ in the Bernoulli stochastic growth models. Therefore, these models are considered nested stochastic models. However, the Gompertz stochastic models given in the following section are not nested with either the logistic or Bernoulli stochastic models.

2.3 Gompertz Stochastic Models

The deterministic Gompertz growth model is given in Table 1. In this paper, we consider the CTMC Gompertz model given by

$$\text{Prob}(\Delta X = j | X(t) = x) = \begin{cases} (ax\log\kappa)\Delta t + o(\Delta t) & j = 1 \\ (ax\log x)\Delta t + o(\Delta t) & j = -1 \\ 1 - (ax\log\kappa + ax\log x)\,\Delta t + o(\Delta t) & j = 0 \\ o(\Delta t) & j \neq 1, -1, 0. \end{cases} \tag{12}$$

The Itô SDE model derived from the CTMC model in Equation (12) is given by

$$dX(t) = aX(t)\log\left(\frac{\kappa}{X(t)}\right) + \sqrt{aX(t)\log\left(\kappa X(t)\right)}\ \ dW. \tag{13}$$

The RDE Gompertz growth model we consider is one in which the growth parameter a is a random variable, $A \sim \mathcal{N}(\mu_a, \sigma_a^2)$, and κ is constant,

$$\frac{dx(t;Q)}{dt} = Ax(t;Q)\log\left(\frac{\kappa}{x(t;Q)}\right). \tag{14}$$

2.4 Dynamic Carrying Capacity Stochastic Models

The dynamic carrying capacity model assumes the carrying capacity κ changes with time and the change is proportional to the value of the population at time t. The deterministic model is given in Table 1. We use the following CTMC dynamic carrying capacity model

$$\text{Prob}(\Delta X = i, \Delta K = j | X(t) = x, K(t) = \kappa) = \begin{cases} (ax\log\kappa)\Delta t + o(\Delta t) & (i,j) = (1,0) \\ (ax\log x)\Delta t + o(\Delta t) & (i,j) = (-1,0) \\ bx^{2/3}\Delta t + o(\Delta t) & (i,j) = (0,1) \\ 1 - \left(ax\log\kappa x + bx^{2/3}\right)\Delta t + o(\Delta t) & (i,j) = (0,0) \\ o(\Delta t) & \text{otherwise.} \end{cases} \tag{15}$$

The Itô dynamic carrying capacity SDE model derived from Equation (15) is given by

$$\begin{aligned} dX(t) &= aX(t)\ln\left(\frac{K(t)}{X(t)}\right) + \sqrt{aX(t)\ln\left(K(t)X(t)\right)}\ dW_1 \\ dK(t) &= bX(t)^{2/3} + \sqrt{bX(t)^{2/3}}\ dW_2 \end{aligned} \tag{16}$$

We further assume the dynamic carrying capacity RDE model has one random variable, the growth parameter a, $A \sim \mathcal{N}(\mu_a, \sigma_a^2)$, and b is constant,

$$\begin{aligned} \frac{dx(t;Q)}{dt} &= Ax(t;Q)\ln\left(\frac{\kappa(t;Q)}{x(t;Q)}\right) \\ \frac{d\kappa(t;Q)}{dt} &= bx(t;Q). \end{aligned} \tag{17}$$

2.5 Power Law Model

The power law model assumes continuous growth of the population at a rate proportional to the population. The deterministic model can be found in Table 1 while the CTMC model is given by

$$\text{Prob}(\Delta X = j | X(t) = x) = \begin{cases} ax^\mu \Delta t + o(\Delta t) & j = 1 \\ 1 - ax^\mu \Delta t + o(\Delta t) & j = 0 \\ o(\Delta t) & j \neq 1, 0. \end{cases} \tag{18}$$

The SDE model derived from Equation (18) is given by

$$dX(t) = aX^\mu(t) + \sqrt{aX^\mu(t)}\ dW \tag{19}$$

and the RDE model assumes the growth parameter a is a random variable, $A \sim \mathcal{N}(\mu_a, \sigma_a^2)$, and μ is constant,

$$\frac{dx(t;Q)}{dt} = Ax^\mu(t;Q). \tag{20}$$

2.6 von Bertalanffy Model

The von Bertalanffy model is a model that has been used to model the growth of organisms, animal populations, aquatic life as well as tumor growth [4, 5, 18, 19, 20, 26] where both the growth rate and natural death rate are proportional to the population. The deterministic model is given in Table 1. We use the CTMC given by

$$\text{Prob}(\Delta X = j | X(t) = x) = \begin{cases} ax^\mu \Delta t + o(\Delta t) & j = 1 \\ bx\Delta t + o(\Delta t) & j = -1 \\ 1 - (ax^\mu + bx)\,\Delta t + o(\Delta t) & j = 0 \\ o(\Delta t) & j \neq 1, -1, 0. \end{cases} \tag{21}$$

The SDE model derived from Equation (21) is given by

$$dX(t) = aX^\mu(t) - bX(t) + \sqrt{aX^\mu(t) + bX(t)}\ dW. \tag{22}$$

The RDE model assumes the growth parameter a is a random variable, $A \sim \mathcal{N}(\mu_a, \sigma_a^2)$, while the death rate parameter b and power of the growth rate μ are assumed to be constant,

$$\frac{dx(t;Q)}{dt} = Ax^\mu(t;Q) - bx(t;Q). \tag{23}$$

2.7 Exponential Growth Model

An exponential growth model assumes exponential growth with a rate of growth a. The CTMC model is given by

$$\text{Prob}(\Delta X = j | X(t) = x) = \begin{cases} ax\Delta t + o(\Delta t) & j = 1 \\ 1 - ax\Delta t + o(\Delta t) & j = 0 \\ o(\Delta t) & j \neq 1, 0. \end{cases} \tag{24}$$

The SDE model derived from Equation (24) is given by

$$dX(t) = aX + \sqrt{aX}\ \ dW \tag{25}$$

and the RDE model assumes the growth parameter a is a random variable, $A \sim \mathcal{N}(\mu_a, \sigma_a^2)$,

$$\frac{dx(t;Q)}{dt} = Ax(t;Q). \tag{26}$$

2.8 Exponential-Linear Model

An exponential-linear growth model assumes an exponential growth for a period of time, $t \leq \tau$ and then constant growth after $t = \tau$ where $\tau = \frac{1}{a_0} \ln\left(\frac{a_1}{a_0 x_0}\right)$ depends on the parameters a_0 and a_1 for the exponential growth rate and constant growth rate, respectively, as well as the initial population size x_0. The CTMC model is given by

$$\text{Prob}(\Delta X = j | X(t) = x) = \begin{cases} a_0 x\Delta t + o(\Delta t) & j = 1 \\ 1 - a_0 x\Delta t + o(\Delta t) & j = 0 \\ o(\Delta t) & j \neq 1, 0, \end{cases} \tag{27}$$

if $t \leq \tau$ and given by

$$\text{Prob}(\Delta X = j | X(t) = x) = \begin{cases} a_1\Delta t + o(\Delta t) & j = 1 \\ 1 - a_1\Delta t + o(\Delta t) & j = 0 \\ o(\Delta t) & j \neq 1, 0, \end{cases} \tag{28}$$

if $t > \tau$. The SDE model derived from Equations (27) and (28) is given by

$$\begin{array}{rcll} dX(t) & = & a_0 X + \sqrt{a_0 X}\ \ dW & t \leq \tau \\ dX(t) & = & a_1 + \sqrt{a_1}\ \ dW & t > \tau. \end{array} \tag{29}$$

The RDE model assumes the growth parameter a_0 is a random variable, $A_0 \sim \mathcal{N}(\mu_{a_0}, \sigma_{a_0}^2)$, but the linear growth rate, a_1 is assume to be constant

$$\begin{array}{rcll} \dfrac{dx(t;Q)}{dt} & = & A_0 x(t;Q) & t \leq \tau \\ \dfrac{dx(t;Q)}{dt} & = & a_1 & t \leq \tau. \end{array} \tag{30}$$

3 Approximation of Stochastic Models by Deterministic Systems

In the paper by Banks and Joyner [11], it was demonstrated how one can use deterministic methods to compare two nested stochastic models by approximating a stochastic model by an appropriate deterministic system. The Kurtz limit theorem, originally developed in [23] (see also [17]), justifies the approximation of

a CTMC by a corresponding deterministic system if the population size N is sufficiently large. The theorem is given by

Kurtz Limit Theorem Let $\boldsymbol{C}^N(t)$ be a continuous-time Markov chain. Suppose that $\lim_{N\to\infty} \boldsymbol{C}^N(0) = \boldsymbol{c}_0$ and for any compact set $\Gamma \in \mathbb{R}^n$ there exists a positive constant η_Γ such that

$$|\boldsymbol{g}(\boldsymbol{c}) - \boldsymbol{g}(\hat{\boldsymbol{c}})| \leq \eta_\Gamma |\boldsymbol{c} - \hat{\boldsymbol{c}}|,$$

for $\boldsymbol{c}, \hat{\boldsymbol{c}} \in \Gamma$. Then we have

$$\lim_{N\to\infty} \sup_{t\leq t_f} |\boldsymbol{C}^N(t) - \boldsymbol{c}(t)| = 0 \tag{31}$$

almost surely for all $t_f > 0$, where $\boldsymbol{c}$ denotes the unique solution to the system of ordinary differential equations given by

$$\dot{\boldsymbol{c}}(t) = \boldsymbol{g}(\boldsymbol{c}), \quad \boldsymbol{c}(0) = \boldsymbol{c}_0.$$

In [24], Ortiz et. al. used this concept as a methodology for estimating parameters of the CTMC model by first approximating the model with its deterministic counterpart and then applying parameter estimation procedures for a deterministic system. Joyner et. al. [21] further tested this methodology and determined that if the population size is 'sufficiently large' (the concept of 'sufficiently large' is model specific), this parameter estimation method produces good estimates. Using the deterministic approximation, Banks and Joyner explained in [11] how one could extend this concept to model comparison for nested stochastic models. They first approximated the CTMC by an appropriate deterministic system. They then used the theory for model comparison of deterministic systems [10] to compare nested CTMC models.

This approach can also be applied to SDEs, since the expectation of the stochastic model is given by the deterministic model. Recall the general form of an SDE given in Equation (4); taking the expectation of the SDE, we have

$$\mathbb{E}\left(dX(t,\mathbf{q})\right) = \mathbb{E}\left(\mu(t,X(t,\mathbf{q}))dt\right) + \mathbb{E}\left(B(t,X(t,\mathbf{q}))dW(t)\right) = \mu(t,X(t,\mathbf{q}))dt$$

or

$$\frac{\mathbb{E}\left(dX(t,\mathbf{q})\right)}{dt} = \mu(t,X(t,\mathbf{q}))$$

since $\mathbb{E}(dW) = 0$. Therefore, the expected trend for an SDE is given by the expected deterministic system.

Recall that a general random ordinary differential equation (RDE) containing random parameter values can be written as in Equation (5), given by

$$\frac{d\mathbf{x}}{dt} = g(t,\mathbf{x},\mathbf{Q}), \quad \mathbf{x}(0) = \mathbf{x}_0,$$

where $\mathbf{Q}$ is a m-dimensional random vector. There are two common ways to approach random differential equations, the mean calculus approach and the sample function approach [7]. Using the sample function approach, one considers *each realization* of the random differential equation to be a deterministic differential equation, called a sample deterministic differential equation, assumed to have a unique solution [7]. For example, in the RDE logistic model, given in Section 2.1, for every realization r of $R \sim \mathcal{N}(\mu_R, \sigma_R^2)$ in Equation (6), we obtain the deterministic differential equation given by Equation (1). In this approach to RDE models, the solution to a random differential equation is a collection of solution trajectories to the sample deterministic equations. Hence, *in each of these stochastic models*, we can approximate the chosen stochastic model by the deterministic system. In the next section, we will discuss details of this approximation and then how to use the Akaike Information Criterion (AIC) for model comparison.

4 Extension of Model Selection Based on AIC for Stochastic Models

For proof of concept, for each of the stochastic growth models given in Sections 2.1 - 2.8, we first simulate data and then attempt to recover the appropriate model when performing a model comparison test. As discussed in the previous section, each type of stochastic model can be approximated by an appropriate deterministic system. Therefore, to compare candidate stochastic models, we first approximate each of the stochastic models with the appropriate deterministic model from Table 1. We then use a deterministic methodology, specifically the Akaike Information Criterion (AIC), to compare the set of approximated deterministic models and assume that the model selection process for the deterministic systems can provide meaningful insight into the "best" model given a set of candidate stochastic models.

AIC is one of the most widely used methods for choosing a "best approximating" model from several competing models given a particular data set [14, 16]. The well-known Akaike Information Criterion is given by

$$AIC = -2\ln\left(\mathcal{L}(\hat{\boldsymbol{\theta}}_{MLE}|\mathbf{y})\right) + 2\kappa_\theta \tag{32}$$

where $\hat{\boldsymbol{\theta}}_{MLE}$ is the maximum likelihood estimate and κ_θ is the total number of parameters in the model; we note that κ_θ includes the number of required parameters for both the model and the assumed distribution. The complexity of the model, as given by the total number of parameters in the model, is considered in the criterion where, given the same level of accuracy, the simpler model is preferable to the more complex one.

In this paper, we consider the AIC under the frameworks of least squares estimation as given in [8]. Depending on the statistical model for the system, the appropriate parameter estimation technique varies. This results in different formulations for the AIC as given below where κ_q is the total number of parameters in the model only, $\mathbf{q}_0$ is the "true" parameter and $\hat{\mathbf{q}}$ is the estimate for $\mathbf{q}_0$ using the appropriate least squares methodology (see [7] for a detailed discussion on choosing the appropriate methodology).

- If the statistical model is assumed to have *absolute error* given by

$$Y_j = f(t_j, \mathbf{q}_0) + \mathcal{E}_j, \tag{33}$$

  we consider an ordinary least squares (OLS) formulation of the AIC given by

$$\mathrm{AIC}_{OLS} = N\ln\left(\frac{\sum_{j=1}^{N}(y_j - f(t_j, \hat{\mathbf{q}}_{OLS}))^2}{N}\right) + 2(\kappa_{\mathbf{q}} + 1). \tag{34}$$

- If the statistical model is assumed to have *constant weighted error* given by

$$Y_j = f(t_j, \mathbf{q}_0) + w_j\mathcal{E}_j, \tag{35}$$

  we consider a weighted least squares (WLS) formulation of AIC given by

$$\mathrm{AIC}_{WLS} = N\ln\left(\frac{\sum_{j=1}^{N} w_j^{-2}(y_j - f(t_j, \hat{\mathbf{q}}_{WLS}))^2}{N}\right) + 2(\kappa_{\mathbf{q}} + 1). \tag{36}$$

- If the statistical model is assumed to have *parameter dependent weighted error* given by

$$Y_j = f(t_j, \mathbf{q}_0) + f^{\gamma}(t_j, \mathbf{q}_0)\mathcal{E}_j, \tag{37}$$

we use the iterative reweighted, weighted least squares (IRWLS) formulation of AIC given by

$$\mathrm{AIC}_{IRWLS} = N \ln \left(\frac{\sum_{j=1}^{N} w_j^{-2}(y_j - f(t_j, \hat{\mathbf{q}}_{IRWLS}))^2}{N} \right) + 2(\kappa_{\mathbf{q}} + 1), \tag{38}$$

where $w_j = f^{\gamma}(t_j; \hat{\mathbf{q}}_{IRWLS})$.

Therefore, the associated statistical model, in addition to the appropriate deterministic system, plays an important role in the formulation of the AIC. We first note that by assuming $\gamma = 0$ in the iterative reweighted weighted least squares formulation, the statistical model simplifies to the ordinary least squares formulation. Hence, to determine an appropriate statistical model, one must find an appropriate value of γ. In practice, there are different methods for determining γ. One approach is to initially assume a specific statistical model, i.e., a specific value of γ, and then carry out the appropriate parameter estimation technique. One can then use residual plots, as in [7], to determine whether or not the assumed statistical model is correct. If the statistical model is appropriate, when one plots modified residuals versus t_j, the residual plots will exhibit a random pattern. *Modified residuals* are given by

$$r_j^{mod} = r_j/|y_j - f(t_j; \hat{\mathbf{q}})|^{\gamma} = (y_j - f(t_j; \hat{\mathbf{q}}))/|y_j - f(t_j; \hat{\mathbf{q}})|^{\gamma} \tag{39}$$

where $\hat{\mathbf{q}}$ is the estimate for $\mathbf{q}_0$. On the other hand, if the assumed statistical model is incorrect, the residual plots will exhibit a non-random pattern, such as a fan-shaped pattern. One can iteratively choose varying values for γ and repeat the process until the appropriate statistical model is determined. This process can be extremely computationally and time intensive as it potentially involves solving a parameter estimation problem multiple times.

Another approach for determining the correct statistical model is to use a difference-based estimation method as proposed in [6]. In this method, one uses pseudo-measurement errors to estimate the variance; see [6] for details on the theory and specific implementation. In this paper, we focus on such a difference-based estimation method involving the second-order differencing of the data to calculate the *pseudo measurement errors*, given by

$$\epsilon_j = \hat{\epsilon}_j^{2nd} = \frac{1}{\sqrt{6}}(y_{j-1} - 2y_j + y_{j+1}). \tag{40}$$

The pseudo measurement errors provide a reasonable approximation of the true measurement errors [6]. Therefore,

$$e_j^{mod} = \epsilon_j/|y_j - \epsilon_j|^{\gamma} \tag{41}$$

provides an estimation of the modified residuals discussed in the previous approach. If one then plots e_j^{mod} versus t_j for various values of γ, one can determine which value of γ results in a random pattern, thus indicating an appropriate value of γ for the statistical model. This process is still iterative; however, it does not involve solving a computationally-intensive parameter estimation problem with each iteration. This method has been shown to provide similar results to the previously described modified residuals using method (39) but without the time-intensive calculations.

Using this second-order difference-based estimation procedure, we determine for each data set which statistical model is appropriate. For each of the continuous-time Markov chain models, $\gamma = 0.5$ is the

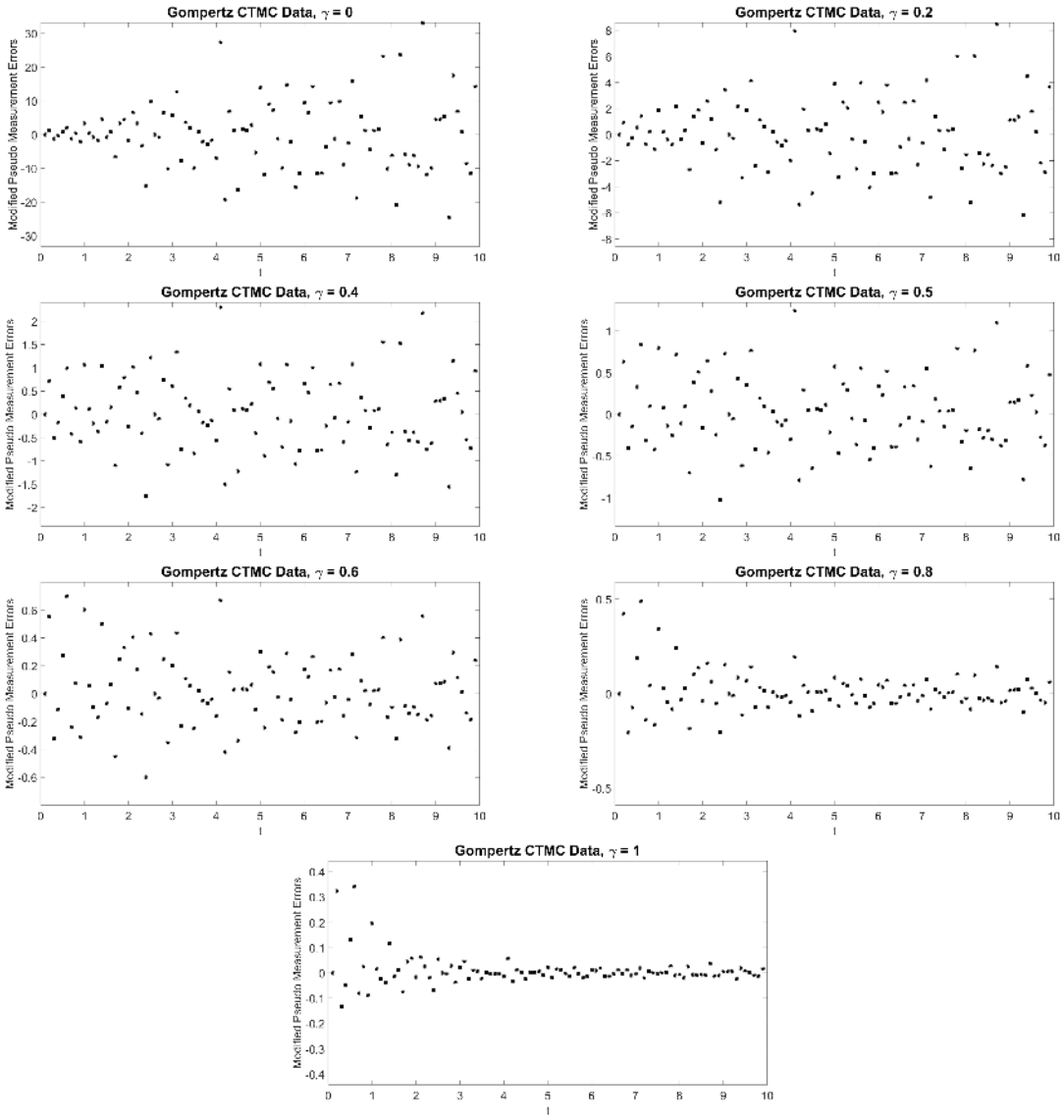

Figure 2: These figures show the modified pseudo-measurement errors (Equation (41)) versus time for simulated Gompertz CTMC data (Equation (12)) with different values of γ where the modified pseudo measurement errors use a second-order difference-based approximation (Equation (40)).

appropriate choice for γ. In Figure 2, we give an example of the modified pseudo measurement errors (41) plotted against time for a simulated data set using the Gompertz CTMC model in Equation (12). For $\gamma = 0$ and $\gamma = 0.2$, there is a fan shape opening to the right where the modified pseudo measurement errors start small and then increase with time. For $\gamma = 0.8$ and $\gamma = 1$, the modified pseudo measurement errors start larger and then decrease as a function of time, producing an inverted fan shape. However, for $\gamma = 0.5$, the modified pseudo measurement errors are randomly distributed. We note that $\gamma = 0.4$ and $\gamma = 0.6$ produce

similar results to $\gamma = 0.5$; however, $\gamma = 0.5$ appears to produce a slightly more randomly distributed pattern across all simulated data sets.

We have a similar result for SDE simulated data. An example of the modified pseudo measurement errors plotted versus time for one of the Gompertz SDE simulated data sets (Equation (13)) is shown for three different values of γ in Figure 3. In this case, $\gamma = 0.6$ produced slightly more randomly distributed points than $\gamma = 0.5$; therefore, for SDE simulated data, we chose $\gamma = 0.6$ in the associated statistical model. We note that the pattern in the pseudo measurement error plots for $\gamma = 0$ and $\gamma = 1$ produce similar shapes as those for CTMC data given previously.

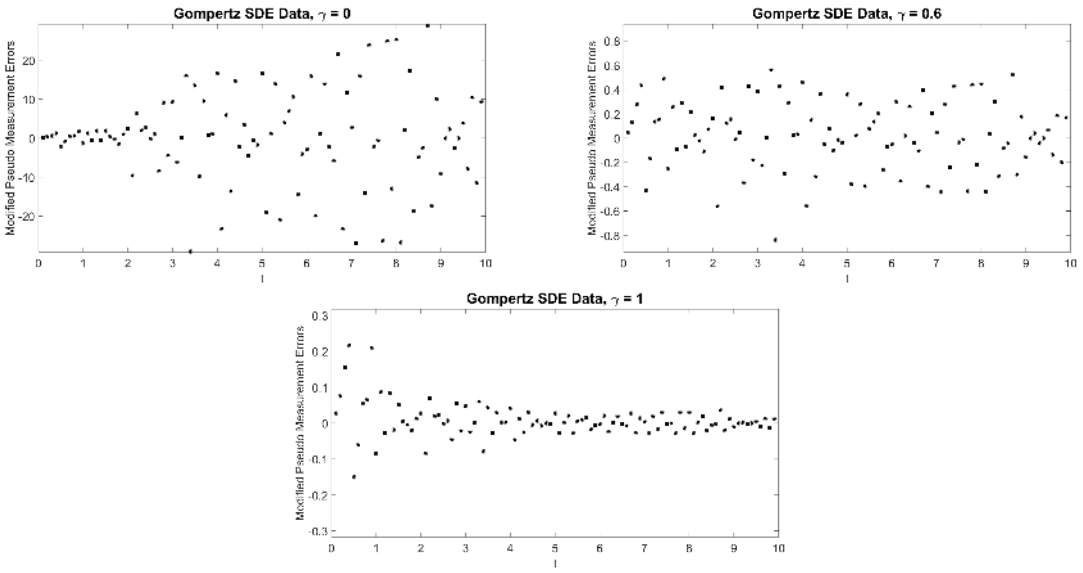

Figure 3: These figures show the modified pseudo-measurement errors (Equation (41)) versus time for simulated Gompertz SDE data (Equation (13)) with different values of γ where the modified pseudo measurement errors use a second-order difference-based approximation (Equation (40)).

For simulated RDE data, we use an ordinary least squares formulation for the statistical mode, i.e., $\gamma = 0$. An example of the modified pseudo measurement error plot versus time for Gompertz RDE data is given in Figure 4. It is clear that using $\gamma = 0$ produces randomly distributed points while for values of $\gamma > 0$, a pattern is apparent in the pseudo measurement error plots.

Given the deterministic approximations to a stochastic model discussed in Section 3 and the appropriate statistical model for the data as discussed above, we propose the following methodology for model comparison of stochastic models and test this methodology in the next section using simulated data.

Stochastic Model Comparison using Deterministic Methodology

Step 1: Approximate each of the candidate stochastic models with an appropriate deterministic model.

Step 2: Choose an appropriate statistical model to describe the data. Using the data and a second-order difference-based procedure as described above, one can determine an appropriate value of γ for the statistical model.

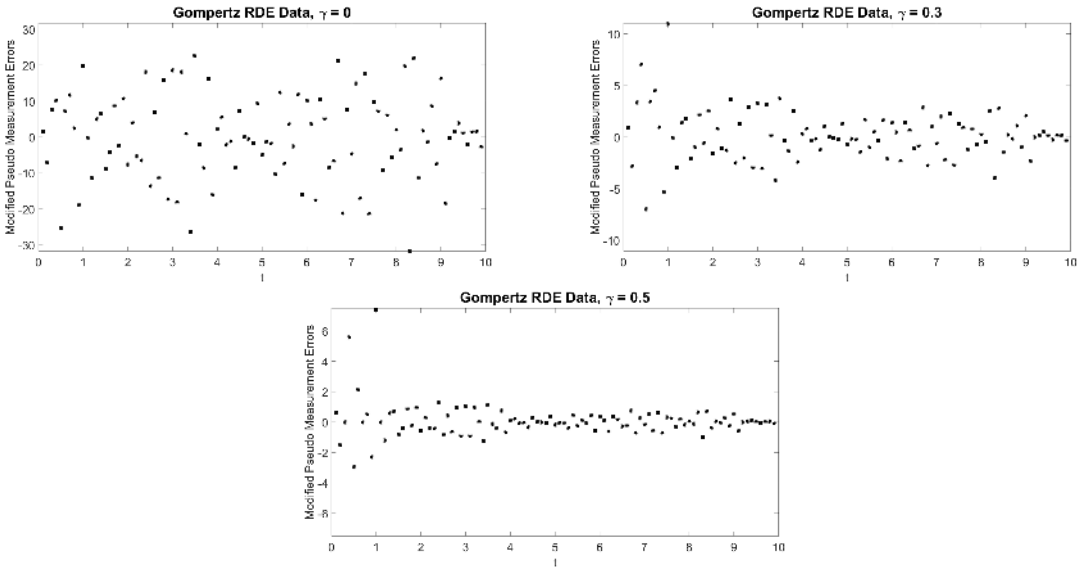

Figure 4: These figures show the modified pseudo-measurement errors (Equation (41)) versus time for simulated Gompertz RDE data (Equation (14)) with different values of γ where the modified pseudo measurement errors use a second-order difference-based approximation (Equation (40)).

Step 3: Use the appropriate methodology for the statistical model chosen in Step 2 to estimate the optimal parameter $\hat{\mathbf{q}}$ for each of the candidate models.

Step 4: Use the appropriate AIC formulation (Equation (34), (36), or (38)) given the statistical model from Step 2, to calculate the AIC value for each candidate model.

Step 5: The "best" deterministic model from the set of candidate models is given by the deterministic model with the lowest AIC value.

Step 6: Given the results of Step 5, information can be obtained about the potential "best" stochastic model from the set of candidate stochastic models.

5 Results using Synthetic Data from Stochastic Models

In this section, we use simulated data from each of the various stochastic growth models to test the accuracy of determining the best candidate stochastic model using the methodology outlined in Section 4. In other words, we first approximate the candidate models with appropriate deterministic models and then use AIC to compare the candidate deterministic model approximations. Given the randomness in each of the data sets, we generate five hundred different data sets for each growth model of each type (e.g. Logistic CTMC model, Logistic SDE model, Logistic RDE model, Bernoulli CTMC model, etc.). We then perform the model comparison test for each of these five hundred data sets of each type and tally how often the AIC value was lowest for each of the different models. The simulated data and results are given for CTMC models in the Section 5.1, SDE models in Section 5.2 and RDE models in Section 5.3.

5.1 Continuous Time Markov Chain Models

In Figure 5, we show the variation in the five hundred different data sets for each of the continuous time Markov chain models. We see similar trends in the data for the logistic growth, Bernoulli growth, Gompertz growth and the dynamic carrying capacity models while the data for the power law, von Bertalanffy, exponential growth and exponential-linear growth exhibit a different growth behavior when compared to the previous models but with similar trends to each other. We will discuss these similarities in more detail when discussing the results of the model comparison.

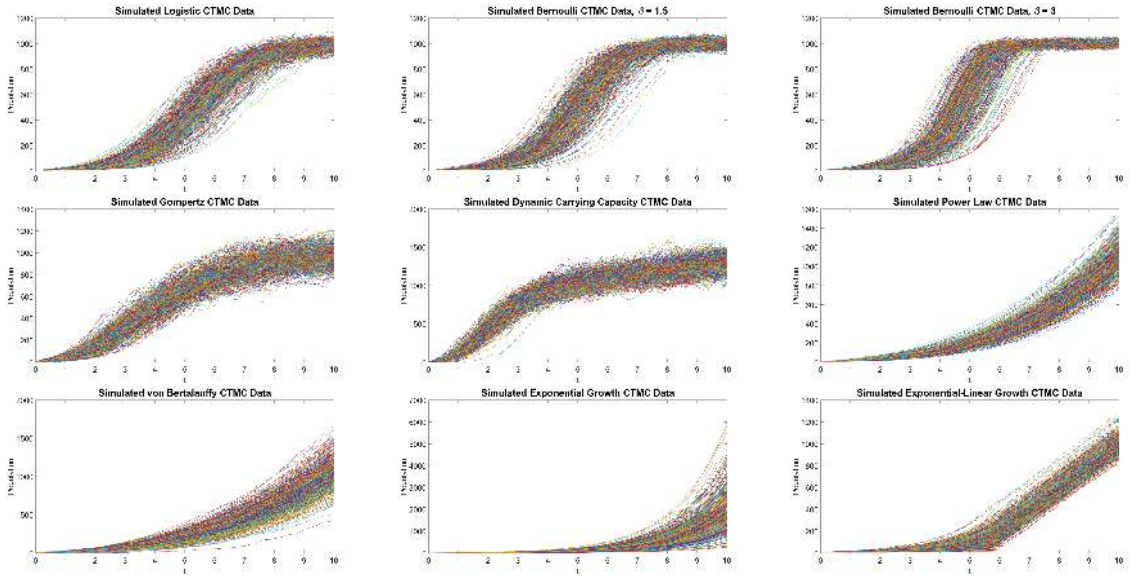

Figure 5: These figures show simulated data for each of the CTMC models: Logistic growth model (Equation (7)), Bernoulli growth model with $\beta = 1.5$ (Equation (9)), Bernoulli growth model with $\beta = 1.5$ (Equation (9)), Gompertz growth model (Equation (12)), Dynamic Carrying Capacity (Equation (15)), Power Law (Equation (18)), von Bertalanffy (Equation (21)), Exponential Growth (Equation (24)), and Exponential-Linear Growth (Equation (28)). The model parameters used in creating the data are the same as those given in Figure 1.

Figure 6 displays the results using the model comparison methodology outlined in Section 4. The bar charts illustrate the frequency with which each model was chosen to be the "best" model from the candidate models for each of the five hundred data sets generated from each growth model. Ideally, if data were generated from a particular stochastic model, e.g., logistic CTMC model, then the model comparison would indicate the generating model, e.g., logistic model, is the "best" model from the candidate models.

Examining the results, we see that for the simulated Bernoulli CTMC data using $\beta = 3$, the generating model (Bernoulli model) was chosen as the best candidate model 100% of the time. In the case in which the generating model was closer to the logistic model with $\beta = 1.5$, the generating model (Bernoulli model) was chosen 94.8% of the time over the other candidate models using the AIC. In the remainder of the cases (5.2%), the logistic model was chosen. Recall that the Bernoulli and logistic models are nested models, i.e., the logistic model can be obtained from the Bernoulli growth model by setting $\beta = 1$. We note that in the few cases in which the logistic model was chosen as the "best" model over the Bernoulli model, the parameter estimates in the Bernoulli model resulted in β estimates close to 1 with a median estimate of $\beta = 1.01$. For

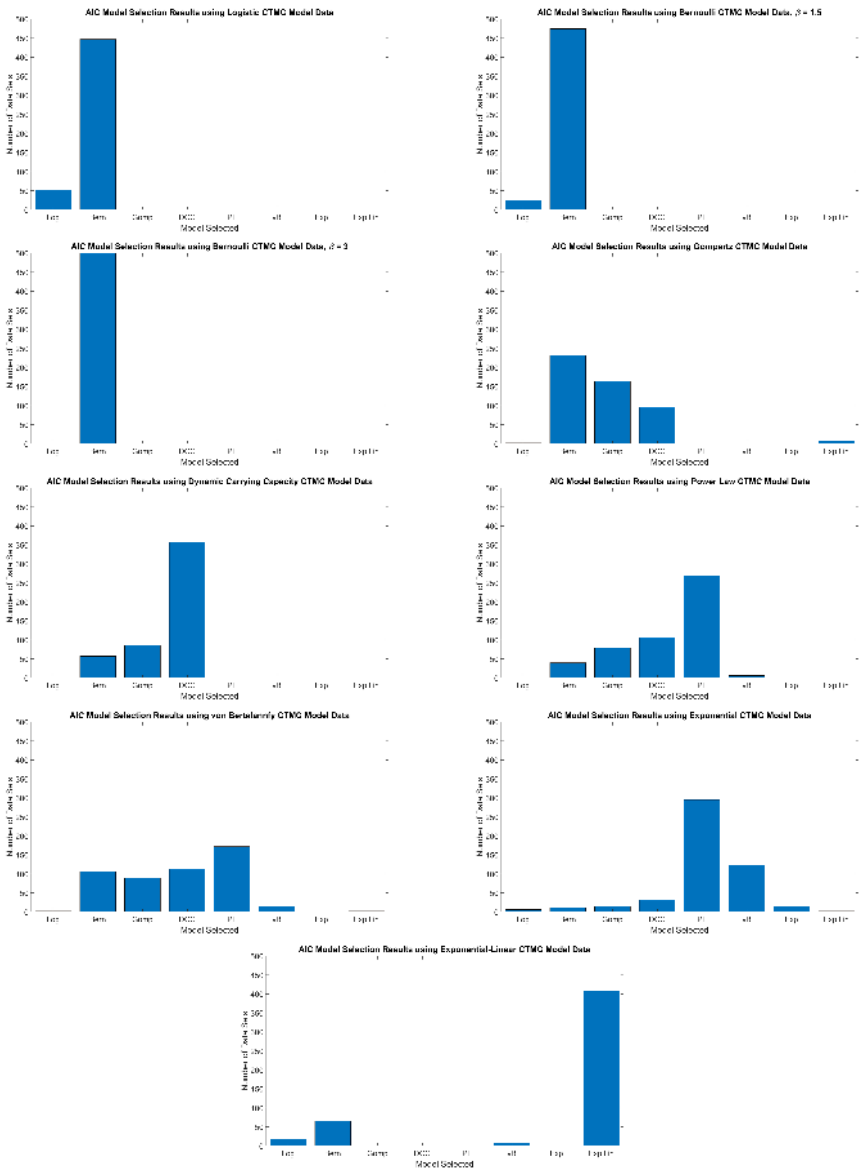

Figure 6: These figures depict the results of the model comparison methodology outlined in Section 4 for the Logistic Growth Model CTMC Data, Bernoulli Growth Model CTMC Data ($\beta = 1.5$), Bernoulli Growth Model CTMC Data ($\beta = 3$), Gompertz Growth Model CTMC Data, Dynamic Carrying Capacity CTMC Data, Power Law CTMC Data, von Bertalanffy CTMC Data, Exponential Growth CTMC Data, and Exponential Linear Growth CTMC Data. For each type of growth model, the "model" from the candidate models as indicated by the AIC values is displayed.

these data sets, the logistic model suffices to explain the data and did not warrant the extra complexity of a third variable in the Bernoulli model.

However, when the simulated data was generated from the logistic CTMC model (the Bernoulli model restricted to $\beta = 1$), the Bernoulli model had the lowest AIC value in 89.4% of the cases; whereas the generating model, the logistic model, resulted in the lowest AIC value in only 10.4% of the cases. Therefore, strictly going by the resulting AIC value, one might conclude that the proposed methodology does not work for stochastic CTMC models. However, the resulting AIC values indicate the "best" candidate model from the *deterministic* model approximations, but one must examine the results further to gain insights into the "best" candidate model from the original stochastic models. In other words, we must also use any relationships between the candidate models, if relationships exist, because when we approximated the original stochastic model with the deterministic approximation, we are only getting an expected behavior of the CTMC model. There will be natural spread about this expected behavior due to the inherent randomness. In the case of the logistic model and Bernoulli model, since they are nested models, we can also examine the resulting estimated values for β (the nested parameter). In the cases in which the Bernoulli model was the chosen model over the logistic model, the median estimated value for β is 1.06 with first and third quartiles given by 0.83 and 1.32 respectively. Hence, by examining the output of the AIC *together* with the relationships between the models and parameter estimates for the models, in all but one case, the proposed methodology pointed to the "best" candidate model being the Bernoulli model with β close to or equal to 1, satisfying the criterion that the data came from a logistic model, or a close logistic model, the Bernoulli model with $\beta \approx 1$. Therefore, the methodology, together with the parameter estimation was successful in choosing the generating model.

If we turn our attention to the simulated data from the Gompertz CTMC model, the results at first seem quite puzzling. In this scenario, the initial AIC results from the deterministic approximations indicate that in 46.4% of the cases the Bernoulli model is the "best" candidate model while only 32.6% of the cases resulted in the actual generating model, the Gompertz model, as the "best" from the candidate models. However, the dynamic carrying capacity was also considered the "best" approximating model in 19% of the cases. In this case, the Gompertz model and dynamic carrying capacity model are clearly related models (see Table 1), although they are only considered nested models if the initial condition κ_0 is also considered a parameter in the system. In that case the Gompertz model can be obtained from the dynamic carrying capacity model by restricting $b = 0$; in that case $\kappa = \kappa_0$ is a fixed limiting capacity parameter, equivalent to the limiting capacity parameter in the Gompertz model. However, we kept κ_0 in our calculations; therefore, these were NOT nested models in this scenario. Nonetheless, we can still compare the output of the two models since the models are clearly related. Recall, if $b \approx 0$, $\kappa \approx \kappa_0$, and growth parameters a are approximately equal, then the two models describe the same population dynamics. In this scenario, the estimated values for κ in the Gompertz model had a median value of 1048 and first and third quartile values of 1018 and 1093 respectively, which were very close to the initial value $\kappa_0 = 1000$. In addition, the estimated growth rate b in the limiting capacity differential equation for the dynamic carrying capacity model was small with a median estimate of $b = 0.16$ and 1st and 2rd quartile values given as 0.08 and 0.25 respectively. Finally, the estimated population growth rate parameters, the parameter a in both models, were almost identical in both parameter estimation problems. The values in the Gompertz model had a median value of 1.14 compared to a median value of 1.17 in the dynamic carrying capacity model. All these factors together indicate that the two models, dynamic carrying capacity model and the Gompertz growth model, are both describing the exact same dynamics and therefore, equivalent models given the estimated parameters within each of the two models. This is similar to the case for the logistic and Bernoulli case described previously. Therefore, in the 51.6% of the cases in which either the Gompertz model or dynamic carrying capacity model were chosen as the "best" of the candidate models, they were describing the same dynamics indicated by the generating model, the Gompertz model.

However, in the other 46.4% of the cases, the Bernoulli model was chosen as the"best" candidate model. At first glance, one might suspect that it is simply because they have similar dynamics and the Bernoulli

model has an extra degree of freedom. However, the Gompertz model is actually related to the Bernoulli model or generalized Logistic model through a limiting process. As shown in [25], since

$$\lim_{v\to\infty} v\left(1-\left(\frac{P(t)}{\kappa}\right)^{1/v}\right) = -\log\left(\frac{P(t)}{\kappa}\right)$$

the Gompertz model is the limit as $v \to \infty$ of the generalized logistic model

$$\frac{dP}{dt} = vaP(t)\left(1-\left(\frac{P(t)}{\kappa}\right)^{1/v}\right) = rP(t)\left(1-\left(\frac{P(t)}{\kappa}\right)^{\beta}\right).$$

Therefore, as $\beta \to 0$, the generalized logistic model and the Bernoulli model will also behave similarly. In the estimation process, the median value of β in the Bernoulli parameter estimation process is 0.008 with first and third quartile values given as 0 and 0.2 respectively. This is consistent with the limiting relationship between the Bernoulli and Gompertz models.

The relationship between these three models can also describe the results seen when using simulated data from the dynamic carrying capacity model. In this case, in 71.4% of the cases, the dynamic carrying capacity, the generating model, was initially chosen using the deterministic approximation. In 17.2% of the cases, the Gompertz model was chosen while in 11.4% of the cases the Bernoulli model was the model with the lowest AIC value out of the candidate models. However, in each of these cases, we can again show that the estimated parameters give dynamics consistent with the generating dynamic carrying capacity model, i.e., they were essentially equivalent models with the same dynamics in these cases.

When the simulated data is generated using the power law CTMC model, in 54% of the cases, the correct generating model is chosen. In 1.2% of the cases, a nested model, the von Bertalanffy model is chosen. However, in 21% of the cases, the dynamic carrying capacity model is the chosen "best" model while in 15.8% of the cases, the Gompertz model is chosen and in 8% of the cases, the Bernoulli is the "best" overall model to describe the data. As described above, there is a clear relationship between the Gompertz, dynamic carrying capacity and Bernoulli models. Intuitively, in the dynamic carrying capacity model, if the growth rate for the carrying capacity is large while the growth rate for the overall population is much smaller, the power law model and the dynamic carrying capacity model could exhibit similar behaviors. However, there is not a clear, defining relationship between the power law and these other three models as in the other cases. Similar results are seen with simulated data for the von Bertalannfy model. In only 3% of the cases was the von Bertalannfy model chosen while in 34.4% of the cases the nested power law model was the chosen "best" model out of the candidate models. Indeed the growth parameter, a, is much larger than the natural death rate parameter b; therefore, the power law is a good approximating model in this case. There appears then to be a similar dynamical relationship between the power law model and the dynamic carrying capacity model (chosen 22.4% of the cases), the Gompertz model (chosen in 18% of the cases) and the Bernoulli model (chosen in 21.4% of the cases).

In the case of simulated data generated from the exponential-linear model, in 83.4% of the cases, the correct generating model was chosen using the deterministic approximation with the AIC values. In the case of simulated data generated from the exponential model, results indicate that the power law was chosen in 59% of the cases while the von Bertalanffy model was chosen in 24.8% of the cases. The exponential model or exponential-linear models were chosen only in 3% and 4% of the cases respectively. However, the exponential, exponential-linear, power law and von Bertalanffy models are all nested models. The exponential model can be attained by restricting $\mu = 1$ in the power model and by restricting $\mu = 1$ and $b = 0$ in the von Bertalanffy model. This is indeed the dynamics found when these models are chosen as the "best" models out of the candidate models. In the power law model, the median value for μ is given by 1.001 while the 1st quartile

and 3rd quartile are given by 0.98 and 1.04 respectively. Similarly in the von Bertalanffy model, the median value for μ is 0.99 while the 1st and 3rd quartile values are 0.93 and 1, respectively. Furthermore, the value for b is much smaller than the estimated value for a in the von Bertalanffy model indicated the growth term is the dominate term over the natural death term. Again, these dynamics are indicative of the exponential model. Hence, by examining the output of the AIC *together* with the relationships between the models and parameter estimates for the models, in 87.2% of the cases, the proposed methodology pointed to the best candidate model being either the von Bertalanffy model with $\mu \approx 1$ and $a >> b$ or the power law model with $\mu \approx 1$, satisfying the criterion that the data came from an exponential model. Therefore, the methodology, together with the parameter estimation, was successful in choosing the generating model.

5.2 Stochastic Differential Equation Models

In this section, we again implement the model comparison methodology discussed in Section 4 using SDE simulated data. Figure 7 shows the variation in the five hundred different SDE data sets for each of growth models. Simulated data from the SDE models are similar in behavior to that seen in the CTMC data (Figure 5); however, there is slightly more variation and more chance of outliers in the data. Results for model comparison are given in Figure 8.

Results are very similar to those given in the previous section for CTMC models. When the generating model is a logistic stochastic differential equation model, the deterministic methodology indicates the "best" model is the logistic model in only 11.8% of the cases but indicates a preference for the Bernoulli model in 88% of the cases. However, if one couples this deterministic methodology with the parameter output and relationships between the logistic and Bernoulli model, in all 99.8% of these cases, the dynamics indicate a slightly modified logistic model, i.e., either the logistic model or a Bernoulli growth model with $\beta \approx 1$; the median value of β is given by 1.07.

When the generating model in the Bernoulli growth model, in 96.8% of the cases, the Bernoulli growth model has the lowest value of AIC, indicating the "best" model out of the candidate models, when $\beta = 1.5$. There is 100% accuracy using just the deterministic methodology when data is generated using $\beta = 3$ in the SDE model.

When the data is generated using the Gompertz SDE model, it is necessary to use the relationships between the Gompertz, Bernoulli, and dynamic carrying capacity models in addition to the deterministic AIC model selection, as the Bernoulli model is the "best" model from the candidate deterministic models in 50.8% of the cases, the Gompertz model is the "best" model in 28.2% of the cases and the dynamic carrying capacity model is the "best" model in 15.8% of the cases. As with the CTMC data, the median estimated value of β in the Bernoulli model is close to 0 (median value of 0.07), indicating similar population dynamics through the limiting relationship between the two models (as discussed in the previous section). In the dynamic carrying capacity model, the carrying capacity growth parameter is again close to 0, indicating an almost constant value of κ_0 which is close to the estimated value of κ in the Gompertz model. Therefore, using both the deterministic approximations together with the dynamics and relationships between the models, the correct generating model is chosen in a majority of the cases.

This relationship between these models shows up again when using simulated data from the dynamic carrying capacity SDE model in which the correct generating model is initially chosen in 74% of the cases. However, one can use the relationship between the Gompertz and dynamic carrying capacity to give an indication for the 15.8% of the cases which are initially classified as originating from a Gompertz model.

When the generating model is the power law SDE model, in approximately 73.2% of the cases, the AIC values using the initial deterministic approximation results in the correct choice of models. However this intuitive relationship between the power law and the other models does come into play in approximately 22% of the cases. In the von Bertalanffy SDE data simulations, the growth factor has a greater impact in

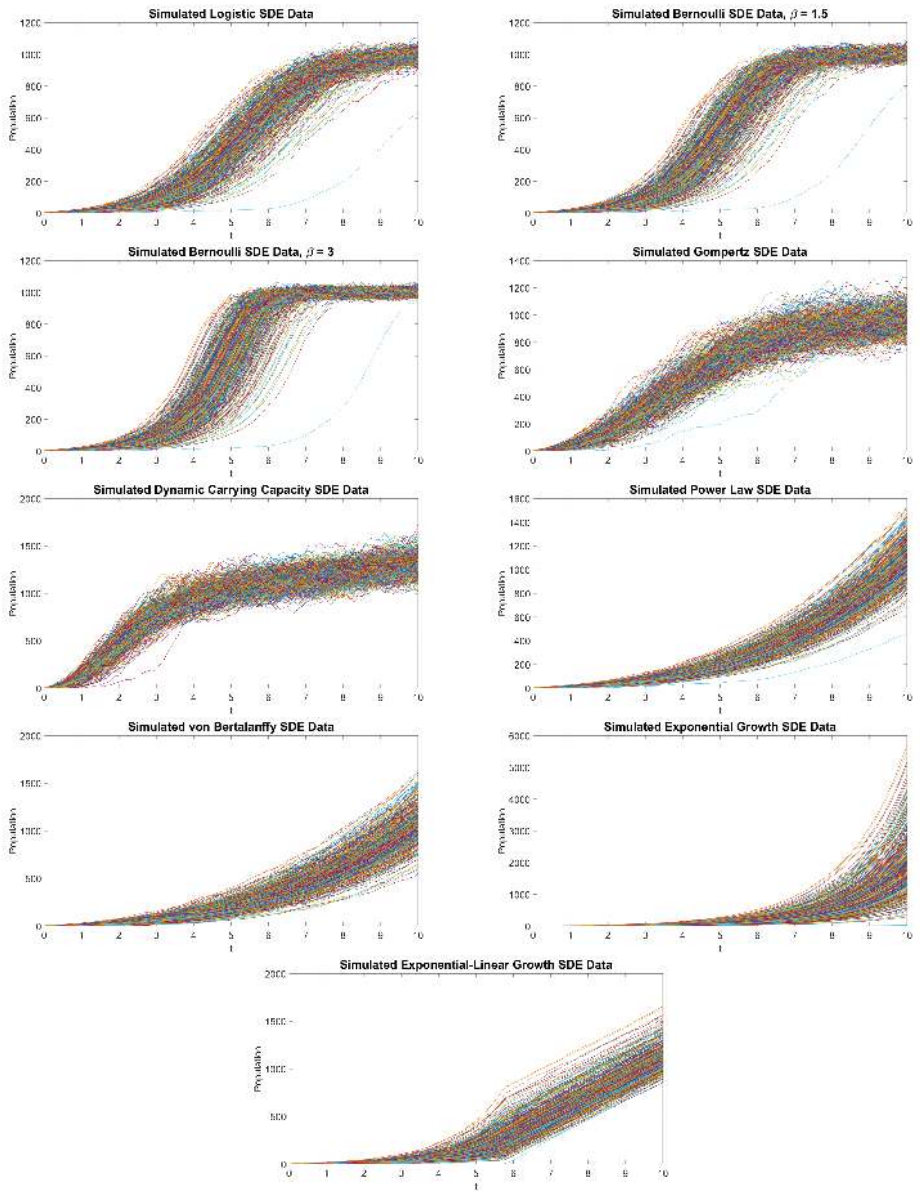

Figure 7: These figures show simulated data for each of the SDE models: Logistic growth model (Equation (8)), Bernoulli growth model with $\beta = 1.5$ (Equation (10)), Bernoulli growth model with $\beta = 1.5$ (Equation (10)), Gompertz growth model (Equation (13)), Dynamic Carrying Capacity (Equation (16)), Power Law (Equation (19)), von Bertalanffy (Equation (22)), Exponential Growth (Equation (25)), and Exponential-Linear Growth (Equation (29)). The model parameters used in creating the data are the same as those given in Figure 1.

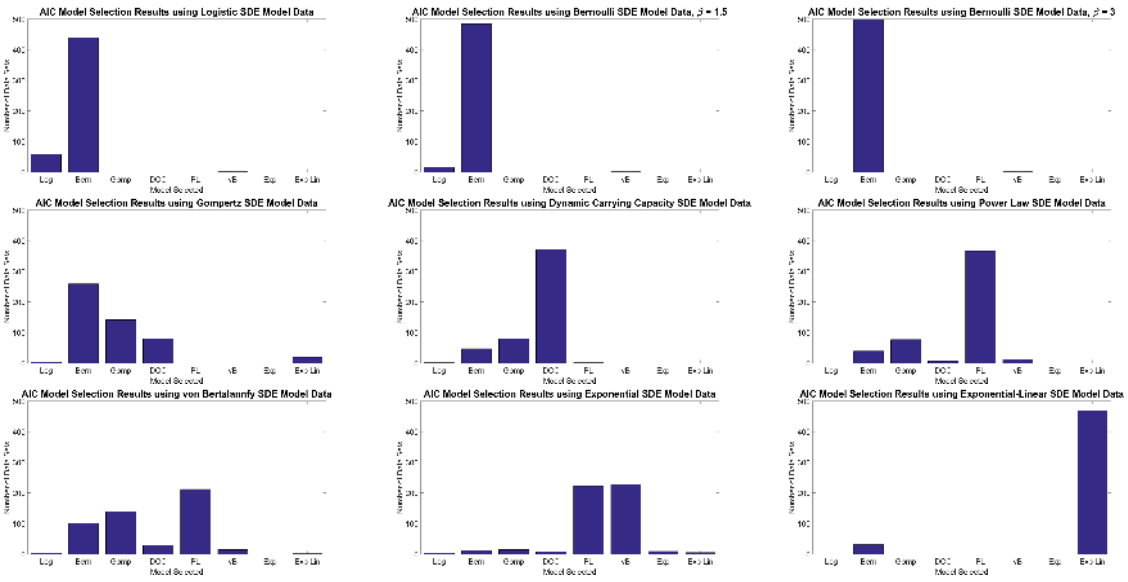

Figure 8: These figures depict the results of the model comparison methodology outlined in Section 4 for the Logistic Growth Model SDE Data, Bernoulli Growth Model SDE Data ($\beta = 1.5$), Bernoulli Growth Model SDE Data ($\beta = 3$), Gompertz Growth Model SDE Data, Dynamic Carrying Capacity SDE Data, Power Law SDE Data, von Bertalanffy SDE Data, Exponential Growth SDE Data, and Exponential Linear Growth SDE Data. For each type of growth model, the "model" from the candidate models as indicated by the AIC values is displayed.

the dynamics than the natural birth rate causing and initial "best" deterministic model to be the nested power law model in 42.4% of the cases.

When using exponential data, the nested relationship between this model and the power law and von Bertalanffy models come into play with a clear indication of the exponential model when examining the parameter estimates with each of these systems. We note that the deterministic methodology indicates the best model is divided almost equally between the power law and the von Bertalanffy models, totally 99% of the cases. Finally, the deterministic methodology alone is able to predict the correct generating model in almost 94% of the cases for data simulated from the exponential linear SDE model.

Therefore, once again, we have shown that the methodology outlined in Section 4 is a feasible method to compare stochastic differential equation models when used in conjunction with the model relationships and parameter estimations as well.

5.3 Random Differential Equation Models

We finally test the methodology on simulated data from random differential equation models as shown in Figure 9. The results of the model comparison are given in Figure 10. Unlike the other two types of stochastic models, when using the sample function approach to RDEs, the solutions are collections of differential equation solutions. Therefore, the deterministic methodology works extremely well in this case without much need for additional insight into the parameters or relationships between the models. In all but one model, the majority of the cases resulted in the correct originating model being chosen as the "best" model from the candidate models. The one exception in when using von Bertalanffy data. In this case, the

growth parameter still overpowers the natural death parameter and the power law is the chosen model in the majority of cases. In almost all other instances, except when using the data from the Gompertz RDE model, the correct model was chosen in almost 70% or more of the cases.

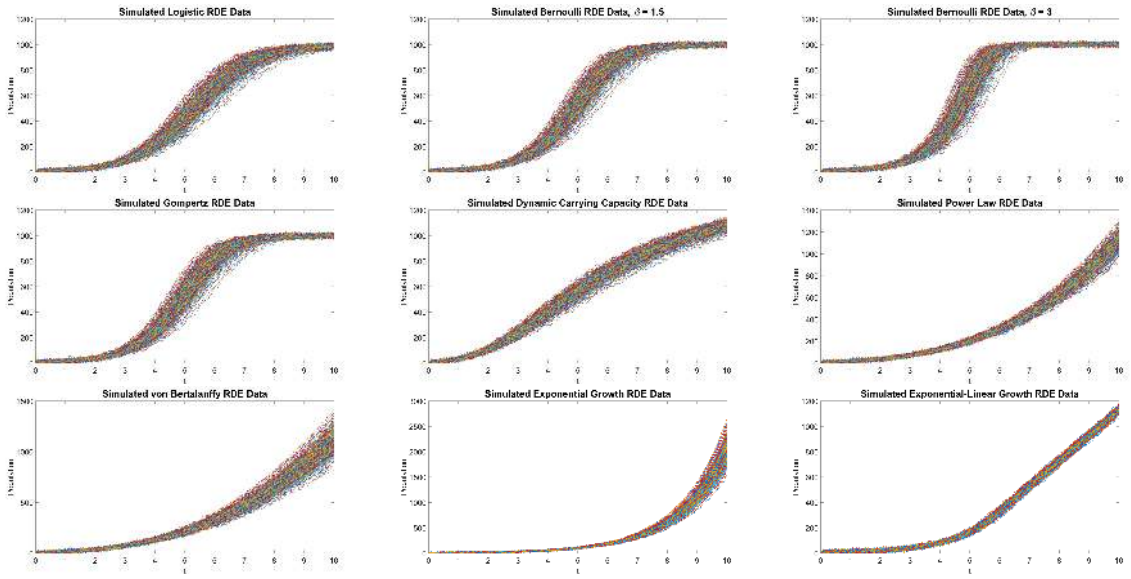

Figure 9: These figures show simulated data for each of the RDE models: Logistic growth model (Equation (6)), Bernoulli growth model with $\beta = 1.5$ (Equation (11)), Bernoulli growth model with $\beta = 1.5$ (Equation (11)), Gompertz growth model (Equation (14)), Dynamic Carrying Capacity (Equation (17)), Power Law (Equation (20)), von Bertalanffy (Equation (23)), Exponential Growth (Equation (26)), and Exponential-Linear Growth (Equation (30)). The model parameters used in creating the data are the same as those given in Figure 1.

6 Results using Experimental Data

To this point, we have developed a model comparison methodology for use on three different types of stochastic models: continuous time Markov chain models, stochastic differential equation models and random differential equation models. For all three types of stochastic models, we first approximated the stochastic model with an appropriate deterministic model. We then determined the value of γ for the statistical model by using a second-order differencing technique [6]. Using the appropriate parameter estimation method for the given statistical model, we estimated the parameters for each of the candidate models and then compared the AIC value for each. We tested this model comparison method using synthetic data from each type of stochastic model for eight different growth systems and illustrated the effectiveness of the technique when using the results of the deterministic AIC value together with information about the growth models themselves and the parameter estimates. In this section, we use the model comparison methodology on longitudinal data collected from algae growth.

In a paper by Banks et. al. [9], longitudinal data was collected from four replicate population experiments with green algae, formally known as *Raphidocelis subcapitata*. Four beakers were initially seeded with 1L of Bold's Basal Media (BBM) and then conditions were set to maintain a chemostat steady-state equilibrium

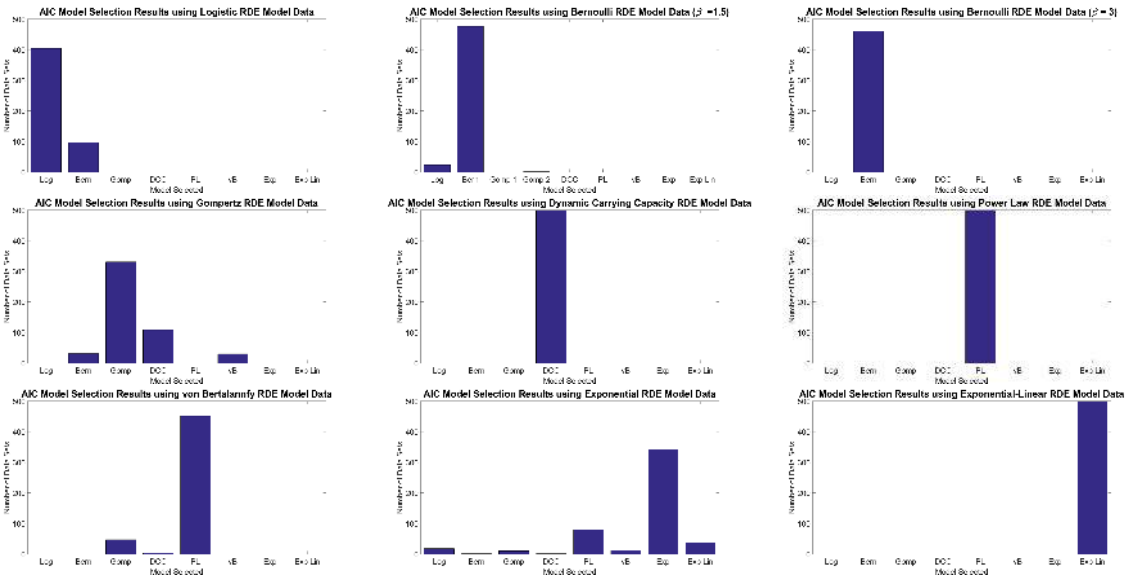

Figure 10: These figures depict the results of the model comparison methodology outlined in Section 4 for the Logistic Growth Model RDE Data, Bernoulli Growth Model RDE Data ($\beta = 1.5$), Bernoulli Growth Model RDE Data ($\beta = 3$), Gompertz Growth Model RDE Data, Dynamic Carrying Capacity RDE Data, Power Law RDE Data, von Bertalanffy RDE Data, Exponential Growth RDE Data, and Exponential Linear Growth RDE Data. For each type of growth model, the "model" from the candidate models as indicated by the AIC values is displayed.

system, constant volume, sufficient oxygen supply, and homogeneous state; details on the experimental collection process can be found in [9]. Two measurements for each of the four replicates were taken twice a day at 9 am and 5 pm daily which were averaged to minimize human measurement error for a total of 36 data points. The data is depicted in Figure 11.

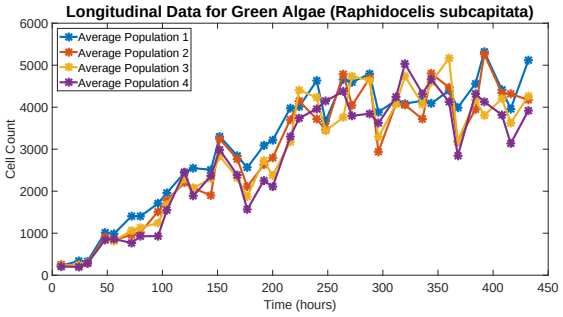

Figure 11: This figure shows the averaged longitudinal data from two measurements each of four replicates of *Raphidocelis subcapitata*.

We assume the data comes from one of the twenty four different stochastic growth models discussed in this paper, CTMC, SDE or RDE logistic, Bernoulli, Gompertz, dynamical carrying capacity, power law,

exponential or exponential-linear growth models. Note that regardless of which *type* of stochastic model we assume is the 'true' model (CTMC, SDE or RDE), we approximate the stochastic model with the same deterministic model (see Table 1). Given an appropriate statistical model for the data, we can then carry out the inverse problem formulation and use the results in the appropriate formulation for AIC. Therefore, the first step is to determine an appropriate statistical model. In Figure 12, we show modified pseudo measurement error plots (Equation (41)) for each of the average population data sets for varying values of γ. There is a fan pattern for both $\gamma = 0$ and $\gamma = 1.4$. We choose $\gamma = 0.8$ for our statistical model.

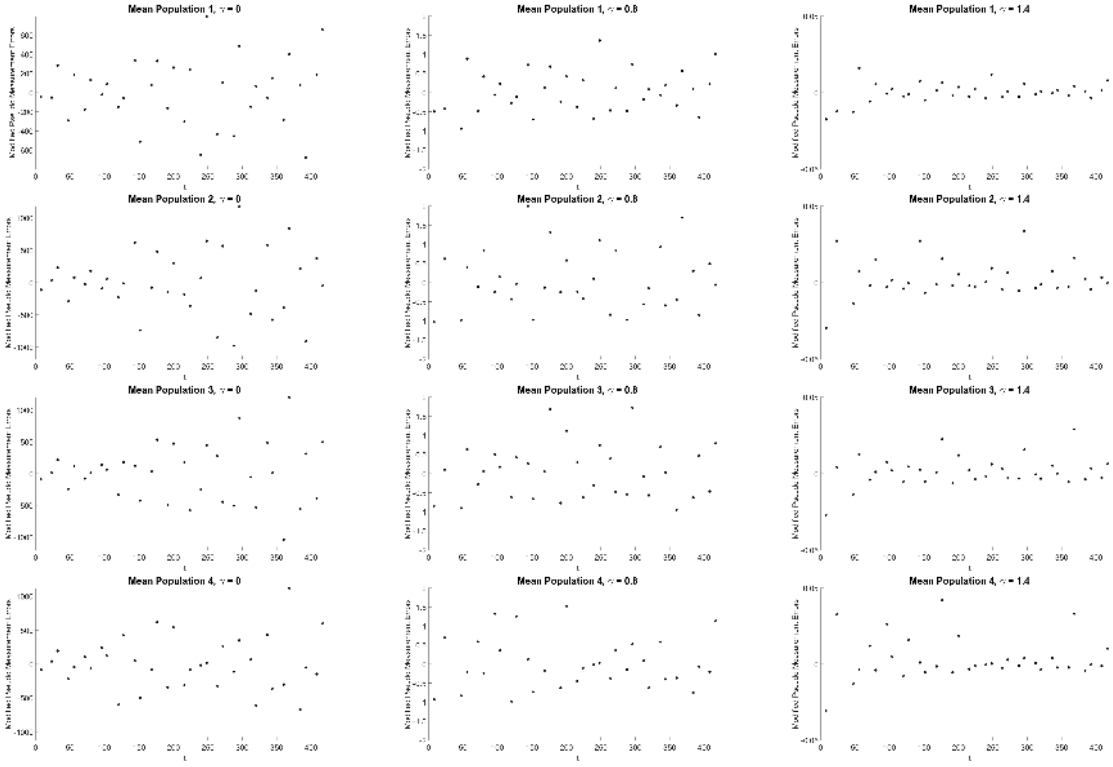

Figure 12: These figures show the modified pseudo-measurement errors (Equation (41)) versus time for the algae population data (Figure 11) with different values of γ where the modified pseudo measurement errors use a second-order difference-based approximation (Equation (40)).

We also must modify the formulation for AIC for this data. In the original formulation of AIC, it is assumed that the sample size is sufficiently large; thus, if the sample size is not sufficiently large relative to the number of parameters which must be estimated, the AIC may perform poorly. In [16], it was suggested that the AIC only be used if the sample size N is at least 40 times as large as the total number of estimated parameters. For small sample sizes, we must modify AIC to account for the small sample size, AIC_c. We use the least squares formulation of AIC_c as given in [8],

$$\text{AIC}_c = \text{AIC} + \frac{2(\kappa_q + 1)(\kappa_q + 2)}{N - \kappa_q}$$

where κ_q are the number of parameters in the model and N is the total number of data points.

Table 2: **AIC_c and Weights for Model Comparison Results for Algae Data**

Pop Data		Log	Bern	Gomp	Dyn Carr Cap	Power	von Bert	Exp	Exp Lin
1	AIC	251.61	243.61	**241.65**	300.55	285.76	291.97	373.26	285.76
	w_i	0.005	0.271	0.724	0	0	0	0	0
2	AIC	264.49	264.71	**263.81**	300.04	286.36	343.26	361.73	286.36
	w_i	0.303	0.272	0.426	0	0	0	0	0
3	AIC	266.85	266.91	**266.23**	302.38	289.30	348.96	364.26	289.30
	w_i	0.300	0.291	0.409	0	0	0	0	0
4	AIC	**272.91**	273.80	273.57	305.85	294.22	343.68	365.37	294.22
	w_i	0.424	0.272	0.304	0	0	0	0	0

We give the results in Table 2 with the lowest AIC highlighted in **red**. We note that when comparing the deterministic model approximations, in three of the four population sets, the AIC values indicate the Gompertz model is the best model out of the set of candidate models to describe this data. In addition to reporting the AIC value, we also calculate Akaike weights which give a normalized relative likelihood of each model. To define the weights, we first define AIC differences $\Delta_i(\text{AIC})$ [16, 27],

$$\Delta_i(\text{AIC}) = \text{AIC}_i - \text{AIC}_{min},$$

where AIC_{min} denotes the minimum calculated AIC value across all candidate models and the term AIC will refer to AIC_c in this case. Akaike [2] indicates that the likelihood of model i given data set $\mathbf{y}$ is proportional to $\exp\left(-\frac{1}{2}\Delta_i\right)$; therefore, it can be used as an indication of the relative strength of evidence for each candidate model. Normalizing the relative likelihoods, the Akaike weights $w_i(AIC)$ are defined by [8, 16, 27],

$$w_i(\text{AIC}) = \frac{\exp\left(-\frac{1}{2}\Delta_i(\text{AIC})\right)}{\sum_{k=1}^{K}\exp\left(-\frac{1}{2}\Delta_k(\text{AIC})\right)} \tag{42}$$

where K is the number of candidate models (eight in our case). We note that the weights of all candidate models sum to 1, so the weight gives a probability that each model is the "best" model. We observe that except for the first population data set, there are almost equal weights between the logistic, Bernoulli and Gompertz models. In the first population data set, β is estimated to be 0.06 for the Bernoulli model while for the other population data sets, β is between 0.4 and 0.56. Since these values are small, it could further indicate that the Gompertz stochastic model is the "best" model for this data set since these models are related through a limiting process.

We plot a hundred simulations of each of the resulting stochastic Gompertz models, together with the algae data, in Figure 13. In the CTMC and SDE models, we assume mean values for the estimates of a and κ in the model. For the RDE model, we assume both a and κ are random variables, $A \sim \mathcal{N}(\mu_a, \sigma_a^2)$ and $\mathcal{K} \sim \mathcal{N}(\mu_\kappa, \sigma_\kappa^2)$ where μ_a, σ_a, μ_κ and σ_κ are estimated using the methodology outlined in [12]. We note that each of the stochastic models have similar trends when compared to the data. It is unclear, however, which particular type of stochastic model might be "best" to describe the data. We further note that assuming x_0 is a parameter might further change the overall fit for the models.

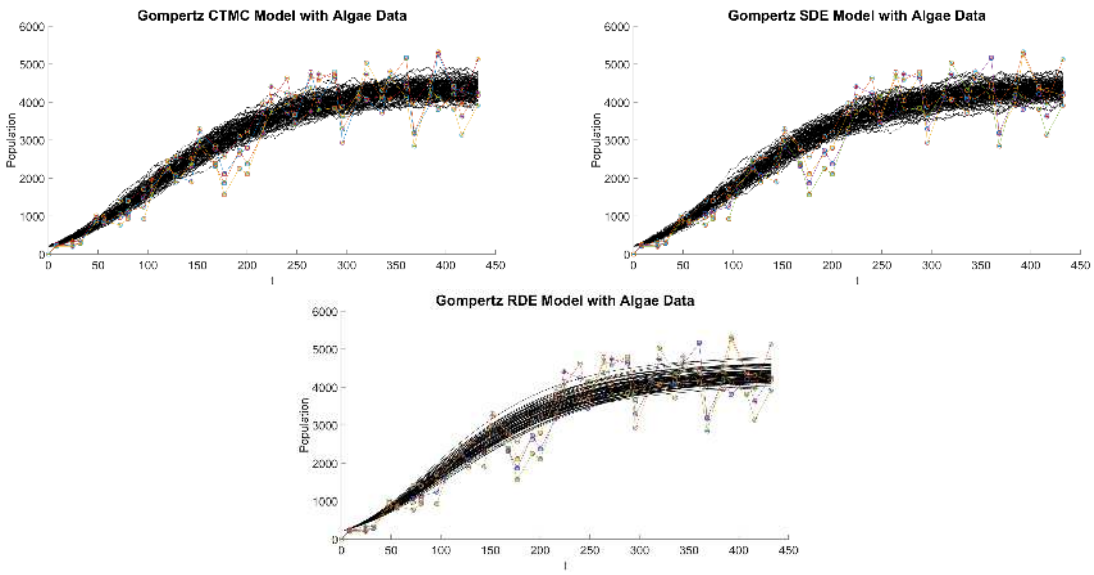

Figure 13: The first plot and second plots show 100 simulations of the Gompertz CTMC and SDE models respectively using the average parameters value estimates of $\overline{\tilde{a}} = 0.0275$ and $\overline{\tilde{\kappa}} = 4428$. The third plot shows 100 simulations of the logistic RDE model when assuming $R \sim \mathcal{N}(\mu_R, \sigma_R)$ and $\mathcal{K} \sim \mathcal{N}(\mu_K, \sigma_K)$ with estimated values $\mu_R = 0.0275$, $\sigma_R = 0.0017$, $\mu_K = 4432$ and $\sigma_K = 170$.

7 Conclusions and Final Remarks

In this paper, we have extended the methodology developed in [11] for comparing two nested stochastic models to a methodology which can be used for an assortment of models, not necessarily nested. Continuous-time Markov chain models, stochastic differential equations and random differential equations can all be approximated by appropriate deterministic models. In this paper, we have shown how one can first approximate a stochastic model by an appropriate deterministic model, together with a correct statistical model given a particular data set, and then use the Akaike Information Criterion for deterministic models to develop insights into the potential best model from the set of candidate stochastic models.

We have illustrated first using simulated data, how this method when used with the resulting parameter estimates and relationships between the candidate models, provides accurate results for the majority of our simulations. We then used this methodology on multiple sets of experimental algae data. Together with the AIC values, AIC weights, parameter estimates and relationships between the models, we illustrated the feasibility of this methodology on experimental data as well. Thus, we have developed a methodology which can be used to discern the "best" model from a set of candidate stochastic models of the same type (for example, a set of candidate CTMC models). In the future, one might investigate how to develop a methodology for determining the "best" *type* of stochastic model, i.e., is the continuous-time Markov chain model, stochastic differential equation model *or* random differential equation model the best *type* of model for the data?

Acknowledgements

This research was supported for both authors in part by the Air Force Office of Scientific Research under grant number AFOSR FA9550-18-1-0457.

References

[1] Kaska Adoteye, H.T. Banks, Karissa Cross, Stephanie Eytcheson, Kevin Flores, Gerald A. LeBlanc, Timothy Nguyen, Chelsea Ross, Emmaline Smith, Michale Stemkovski, and Sarah Stokely. Statistical validation of structured population models for *Daphnia magna*. *Mathematical Biosciences*, 266:73–84, 2015.

[2] Hirotugu Akaike. Information measures and model selection. *Bulletin of the International Statistical Institute*, 50(1):277–291, 1983.

[3] Linda Allen. *An Introduction to Stochastic Processes with Applications to Biology*. Taylor and Francis Group, LLC, 2011.

[4] Libor Babák, P Šupinová, R Burdychová, et al. Growth models of thermus aquaticus and thermus scotoductus. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, 60(5):19–26, 2013.

[5] H. T. Banks, Kidist Bekele-Maxwell, Judith Canner, Amanda Mayhall, Jennifer Menda, and Marcella Noorman. The effect of statistical error model formulation on the fit and selection of mathematical models of tumor growth for small sample sizes. *CRSC-TR17-26, N. C. State University, Raleigh, NC, November, 2017; Int. J. Pure and Applied Mathematics*, 117:203–234, 2017.

[6] H. T. Banks, Jared Catenacci, and Shuhua Hu. Use of difference-based methods to explore statistical and mathematical model discrepancy in inverse problems. *CRSC-TR15-05, N. C. State University, Raleigh, NC, May, 2015; Journal of Inverse and Ill-posed Problems*, 24(4):413–433, 2016.

[7] H. T. Banks, Shuhua Hu, and W. Clayton Thompson. *Modeling and Inverse Problems in the Presence of Uncertainty*. CRC Press, Chapman & Hall, Boca Raton, FL, 2014.

[8] H. T. Banks and Michele L. Joyner. AIC under the framework of least squares estimation. *CRSC-TR17-09, N. C. State University, Raleigh, NC, May, 2017; Applied Mathematics Letters*, 74:33–45, 2017.

[9] H.T. Banks, Elizabeth Collins, Kevin Flores, Prayag Pershad, Michael Stemkovski, and Lyric Stephenson. Standard and proportional error model comparison for logistic growth of green algae (*Raphidocelis subcapiala*). *Applied Mathematical Letters*, 64:213–222, 2017.

[10] H.T. Banks and B.G. Fitzpatrick. Statistical methods for model comparison in parameter estimation problems for distributed systems. *Journal of Mathematical Biology*, 28(5):501–527, 1990.

[11] H.T. Banks and Michele L. Joyner. Determinisitc methodology for comparison of nested stochastic models. *CRSC-TR16-13, NCSU, Raleigh, NC, November 2016; Communication in Applied Analysis*, 21:15–50, 2017.

[12] H.T. Banks and Michele L. Joyner. Parameter estimation for random differential equation models. *CRSC-TR16-15, NCSU, Raleigh, NC, December,2016; Eurasian Journal of Mathematical and Computer Applications(EJMCA)*, 5:5–44, 2017.

[13] H.T. Banks and H.T. Tran. *Mathematical and Experimental Modeling of Physical and Biological Processes*. CRC Press, Chapman & Hall, Boca Raton, FL, 2009.

[14] Hamparsum Bozdogan. Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52(3):345–370, 1987.

[15] Hamparsum Bozdogan. Akaike's Information Criterion and recent developments in information complexity. *Journal of Mathematical Psychology*, 44(1):62–91, 2000.

[16] Kenneth P Burnham and David R Anderson. *Model Selection and Multimodel Inference: A Practical information-Theoretic Approach*. Springer-Verlag, New York, 2nd edition, 2002.

[17] Stewart N. Ethier and Thomas G. Kurtz. *Markov Processes: Characterization and Convergence*, volume 282. John Wiley & Sons, 2009.

[18] Philip Gerlee. The model muddle: in search of tumor growth laws. *Cancer Research*, 73(8):2407–2411, 2013.

[19] Fay Helidoniotis, Malcolm Haddon, Geoff Tuck, and David Tarbath. The relative suitability of the von Bertalanffy, Gompertz and inverse Logistic models for describing growth in blacklip abalone populations (haliotis rubra) in Tasmania, Australia. *Fisheries Research*, 112(1-2):13–21, 2011.

[20] Cynthia M Jones. Fitting growth curves to retrospective size-at-age data. *Fisheries Research*, 46(1-3):123–129, 2000.

[21] Michele Joyner and Thomas Robacker. Development of the MCR method for estimation of parameters in continuous time Markov Chain models. *International J. of Pure and Applied Mathematics*, 112(2):381–416, 2017.

[22] Solomon Kullback and Richard A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[23] Thomas G. Kurtz. Solutions of ordinary differential equations as limits of pure jump Markov processes. *Journal of Applied Probability*, 7(1):49–58, 1970.

[24] A.R. Ortiz, H.T. Banks, Carlos Castillo-Chavez, G. Chowell, and X. Wang. A deterministic methodology for estimation of parameters in dynamic Markov chain models. *Journal of Biological Systems*, 19(01):71–100, 2011.

[25] Michael Stemkovski, Robert Baraldi, Kevin B. Flores, and H.T. Banks. Validation of a mathematical model for green algae (*Raphidocelis subcapitata*) growth and implications for a coupled dynamical system with *Daphnia magna*. *Applied Sciences*, 6(5):155–173, 2016.

[26] Ludwig Von Bertalanffy. Problems of organic growth. *Nature*, 163(4135):156–158, 1949.

[27] Eric-Jan Wagenmakers and Simon Farrell. AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, 11(1):192–196, 2004.